IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 10, Issue 6, June 2021||

DOI:10.15680/IJIRSET.2021.1006022

COVID-19 Pandemic Risk Analysis Using Hyper Parameterized Tuned Machine Learning Model Approach

Jit Dutta[1], M.P Gopinath[2]

P.G. Student, Department of Computer Science, VIT Vellore, Tamilnadu, India[1]

Associate Professor, Department of Computer Science, VIT Vellore, Tamilnadu , India[2]

ABSTRACT: In the healthcare industry, machine learning is commonly used to predict deadly diseases. The aim of this study was to establish and compare the output of a conventional method with a proposed system that uses the Random Forest, K-Nearest Neighbour, Decision tree, Logistic regression, and Naive Bayes classification models to predict COVID-19 disease. The proposed method assisted in tuning the hyper parameters by applying grid search and random search approaches to the five mentioned classification algorithms. The key research subject is the performance of the COVID 19 disease prediction. The hyper parameter tuned model can be used to enhance the efficiency of prediction models. In terms of accuracy, both the conventional and proposed systems were evaluated and compared. The traditional method had accuracy ranging from 43.71 percent to 91.16 percent. The proposed hyper parameter tuning model achieved accuracy levels ranging from 73.31 percent to 94.09 percent. These tests showed that the proposed prediction method is capable of producing more reliable results in predicting COVID-19 disease than the conventional method while maintaining a feasible performance.

KEYWORDS:Covid-19 Disease Prediction, Logistic Regression, K Nearest Neighbour, Machine Learning, Naive Bayes, Decision Tree, Random Forest, Grid Search, Hyper parameter Tuning

I. INTRODUCTION

Machine learning is an evolving topic in the healthcare industry for detecting disease and diagnosis, discovering medications, and classifying medical images. The SARS-CoV-2-caused novel coronavirus disease 2019 (COVID-19) pandemic continues to be a significant and severe threat to the global health. As of March 2021, the estimated number of patients diagnosed with the disease had reached 11,438,734 in more than 180 countries, though the number of people infected is likely much higher. COVID-19 has claimed the lives of 159,079 people [2, 3]. According to WHO reports, the COVID-19 transmission is divided into four stages. The first stage begins with the cases reported by people who travelled through already-affected areas. In the second stage, cases are reported locally among family, friends, and others who came into contact with the person arriving from the affected region. At this stage, the people who have been affected can be identified. The condition becomes much worse in the third level, when the transmission source becomes untraceable and spreads to people who have neither travelled nor come into contact with the infected person. This situation necessitates a nationwide lockdown to eliminate individual social interactions and monitor the rate of transmission. When modern medical technology is unavailable, early diagnosis and prediction of COVID-19 disease are more difficult.

People demand the highest quality of treatment facilities and services in the healthcare sector. The aim of this research is to use a grid and random search to improve the performance of machine learning algorithms. The optimal parameters of the machine learning algorithms can be selected after applying grid and random search. The COVID 19 disease prediction system's efficiency can be improved by tuning these hyper parameters.

The contributions to the research work are listed below.
In the first step, the authors used Random Forest (RF), K nearest neighbor (KNN),

- **Decision tree (DT), Logistic Regression (LR), and Naive Bayes algorithms to create a conventional COVID-19 disease prediction model.**

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 10, Issue 6, June 2021||

DOI:10.15680/IJIRSET.2021.1006022

- **In the second phase, the authors proposed a prediction method, which uses five machine learning algorithms and a hyper parameter tuning approach. To pick the ideal hyper parameters for each, a Grid and Random search is used.**
- **Finally, compared the performance of these two systems in terms of accuracy and precision matrix by the standard state.**

II. RELATED WORK

The contributions of current COVID-19 disease prediction approach are primarily outlined in this section. To predict the COVID 19 disease, researchers have developed a number of machine learning classification models.
The study [4] stated a DT model using hyper parameter tuning that was evaluated on 102 heterogeneous datasets. The research work [5]proposed a machine learning-based computational framework that can be used in emergency to predict COVID-19 patients' risk quickly and accurately. The research work [6]used machine learning algorithms such as logistic regression, decision trees, support vector machines, naive Bayes, and artificial neural networks to create supervised machine learning models for Mexico's COVID-19 infection using epidemiology labelled dataset for positive and negative outcomes.. The research work [8] stated, digital innovations are playing an increasingly significant role in major health-care issues, such as disease prevention. The latest global health emergency, COVID-2019, is also requesting technical assistance. The authors of this paper explored how to use current digital technologies including the internet of things (IOT), big data analytics, artificial intelligence (AI), deep learning, and block chain technology to build strategies for disease surveillance, tracking, and prevention, as well as to assess the effect of the epidemic on the healthcare sector. The paper [9] authors to forecast the distribution of COVID-19, the authors proposed an autoregressive integrated moving average (ARIMA) model. Based on a report on the prevalence and incidence of the COVID-19, the author forecasted various parameters for the next two days in this paper. The correlogram and ARIMA forecast graph for disease occurrence and prevalence are also seen in this work. The paper [10] proposed a time series approach for evaluating the COVID-19 outbreak's incidence pattern and predicted reproduction number They used statistical analysis to look at the patterns of the epidemic in order to illustrate the current epidemiological stage of an area and classify various policies to tackle the COVID-19 pandemic in different countries. The Paper [11] developed a scientific model of crucial SARS-CoV-2 transmission using various datasets to study the COVID-19 outbreak within and outside Wuhan. They used this to look at the likelihood of a disease epidemic spreading outside of Wuhan. The study [12] stated with the available health records data, an Exploratory Data Analysis was conducted on the COVID-19 outburst. This study focused on verified, death, and recovered cases in Wuhan and around the world in order to assess the risk and prepare potential containment activities. The paper [13] stated according to their observations, the criticality of the incubation period for COVID-19 was boosted. They looked at 181 confirmed cases and discovered that the basis duration could range from 5 to 14 days, depending on the outcome.
The researchers developed a covid-19 disease prediction framework that does not require any hyper parameter tuning. So, we've proposed a COVID 19 disease prediction method based on various machine learning algorithms and a hyper parameter tuning approach.

III. METHODOLOGY

In this section, the technical aspects of our approach are discussed. The primary aim of this research is to explore prediction systems in order to design an automated medical diagnosis system that utilizes the collected data.

Figure-1 displays the main flow diagram of the COVID-19 disease prediction method.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 10, Issue 6, June 2021||

DOI:10.15680/IJIRSET.2021.1006022

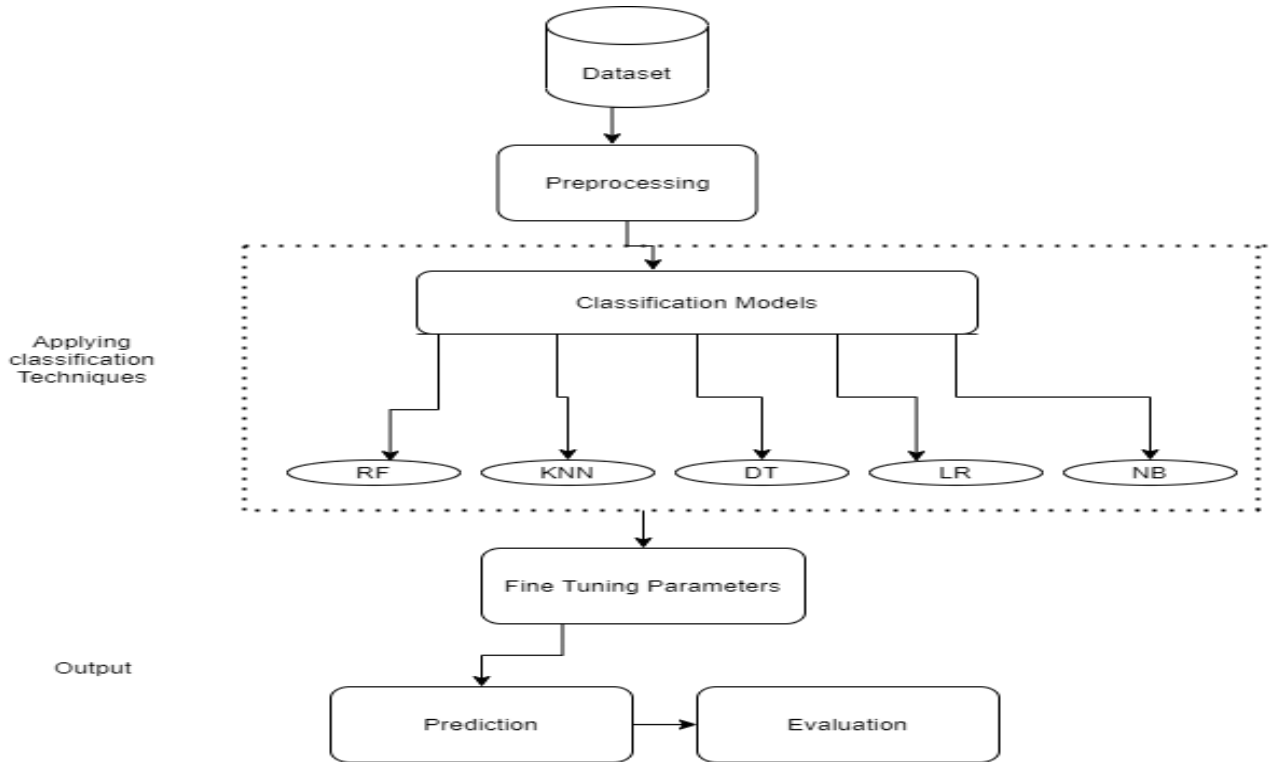


Figure-1

Sub Flow Diagram of the traditional approach with our proposed approach

The study took into account two levels of COVID-19 disease prediction. Figure 2 displays the sub flow diagram of the conventional COVID-19 disease prediction method without a hyper parameter tuning approach to the machine learning algorithm.

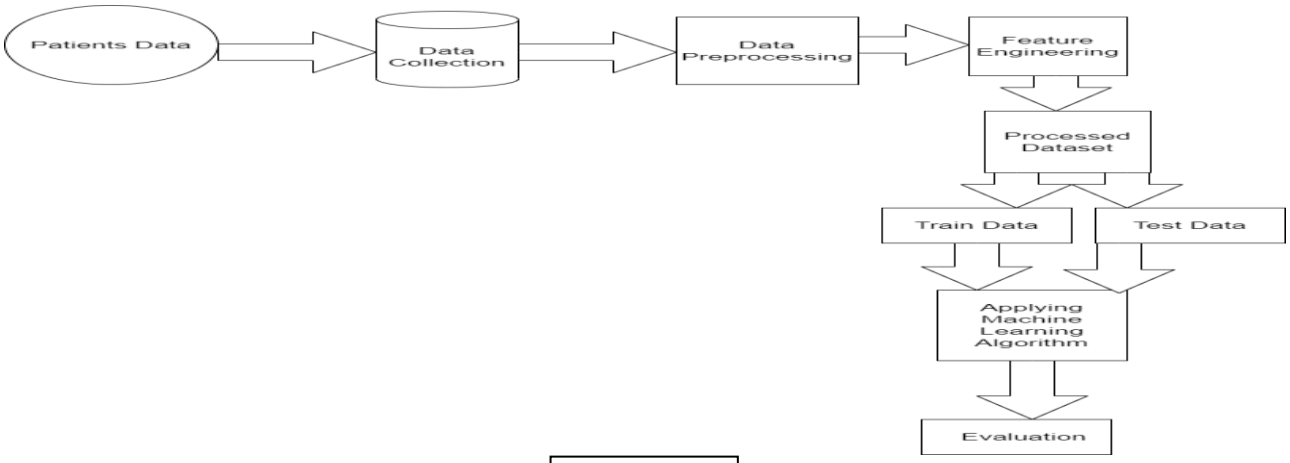


Figure-2

Figure-3 shows the sub flow diagram of a prediction method with a hyper parameter tuning approach to the machine learning algorithm proposed in this report.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 10, Issue 6, June 2021||

DOI:10.15680/IJIRSET.2021.1006022

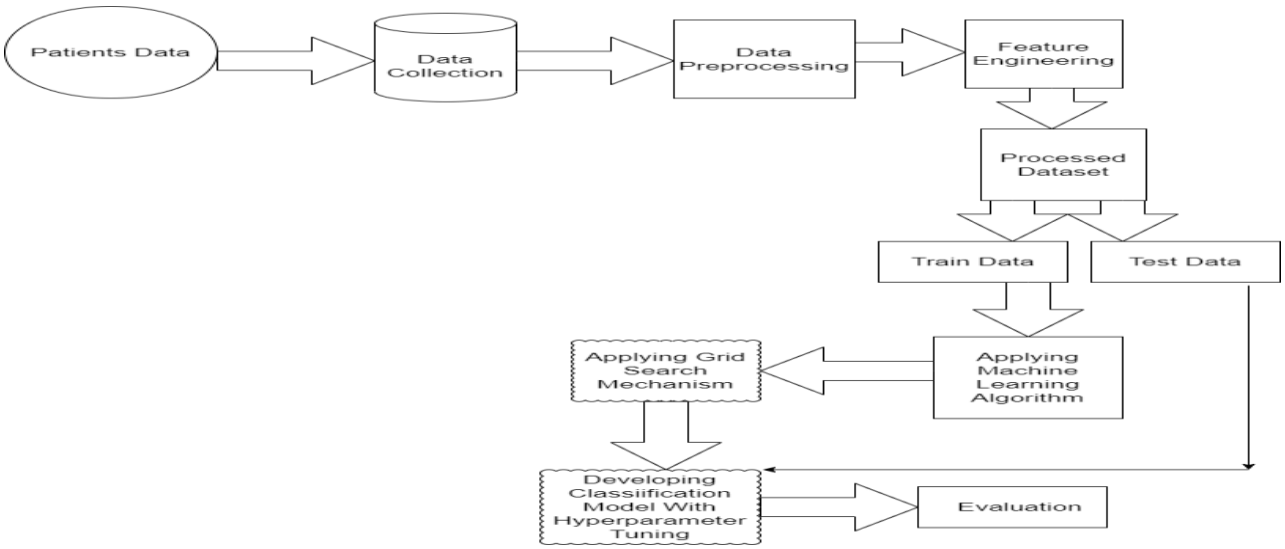


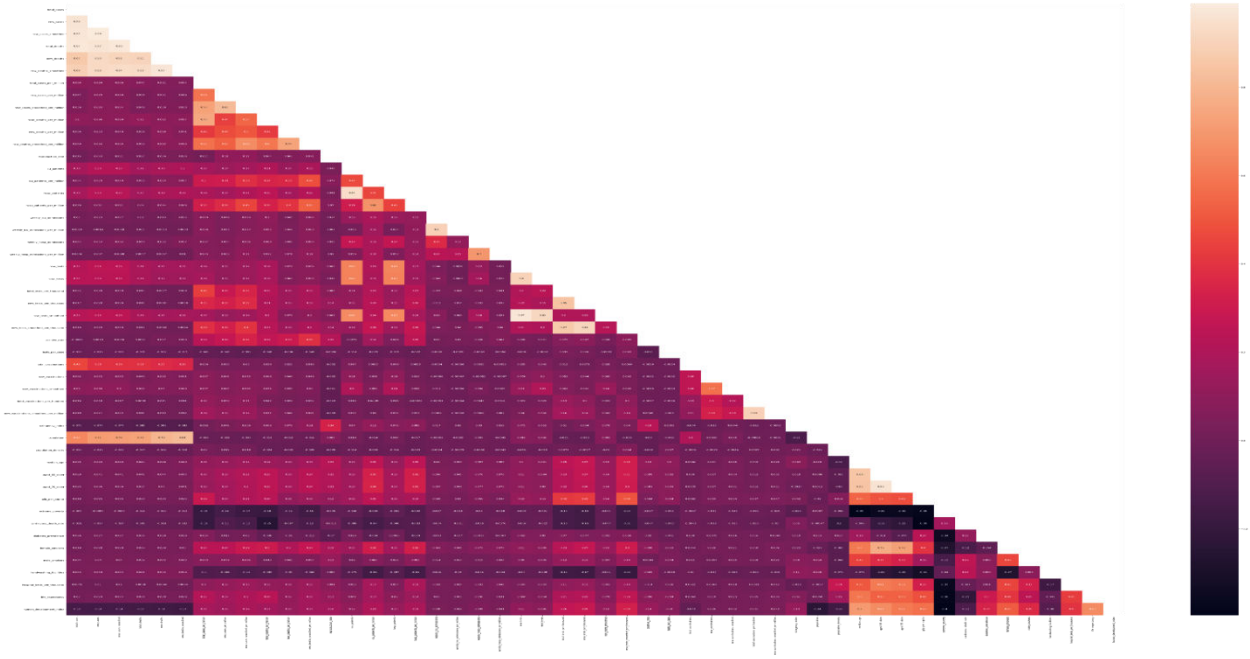
Figure-3

IV. EXPERIMENTAL RESULTS

Result of data pre-processing

There are 56789 samples in the COVID-19 disease dataset, with 60 attributes. The statistical procedure was performed in the pre-processing phase to define and delete missing values and standardize the data.

The heat map in Figure-4 depicts the association between the features of the COVID-19 disease dataset. The values on the two-dimensional surface have been depicted using various colours. Continuous-valued attributes are more concentrated than categorical-valued attributes, as can be observed. The hierarchical clustering and general view of numeric data is depicted by heat map of the COVID-19 disease dataset.



| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 10, Issue 6, June 2021||

DOI:10.15680/IJIRSET.2021.1006022

Experimental results of the traditional and proposed system

The following sections describe the experimental results of various classifiers from the conventional and proposed schemes.

Performance evaluation and comparison of the traditional system

Traditional System	Performance Evaluation Of Test Dataset			
Machine Learning Algorithms	Accuracy	F1	Recall	Precision
RF	91.66	84.75	80.90	88.98
KNN	92.48	86.27	82.24	90.71
DT	76.46	87.88	90.62	89.23
LR	73.35	0	0	0
NB	43.71	41.23	26.30	95.39

Table-1

Performance evaluation and comparison of the proposed system

The Grid search is used in the proposed method to find the best hyper parameters. The classification models are created after the hyper parameters have been fine-tuned. The proposed system's outcome is shown in Table 2,

Proposed System With Tuning		Performance Evaluation On Test Data			
Model	Parameter	Accuracy	F1	Recall	Precision
RF	n_estimators:10, 1000, 50 max_depth:None, 3, 5, 10 min_samples_split:. 2, 20, 2 min_samples_leaf:, 1, 20, 2	94.09	89.80	81.50	100
KNN	No of neighbor= 1,21	93.51	88.69	81.19	97.72
DT	criterion: 'gini' ,'entropy' max_depth: 1,10 min_samples_split: 1,10 min_samples_leaf: 1,5	93.57	88.79	81.29	97.82
LR	C: -4, 4, 50 penalty: 'l1', 'l2' fit_intercept: True, False	74.12	5.01	57.34	2 62
NB	var_smoothing:0,-9, num=100	73.31	8.6	39.9	4.8

Table-2

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 10, Issue 6, June 2021||

DOI:10.15680/IJIRSET.2021.1006022

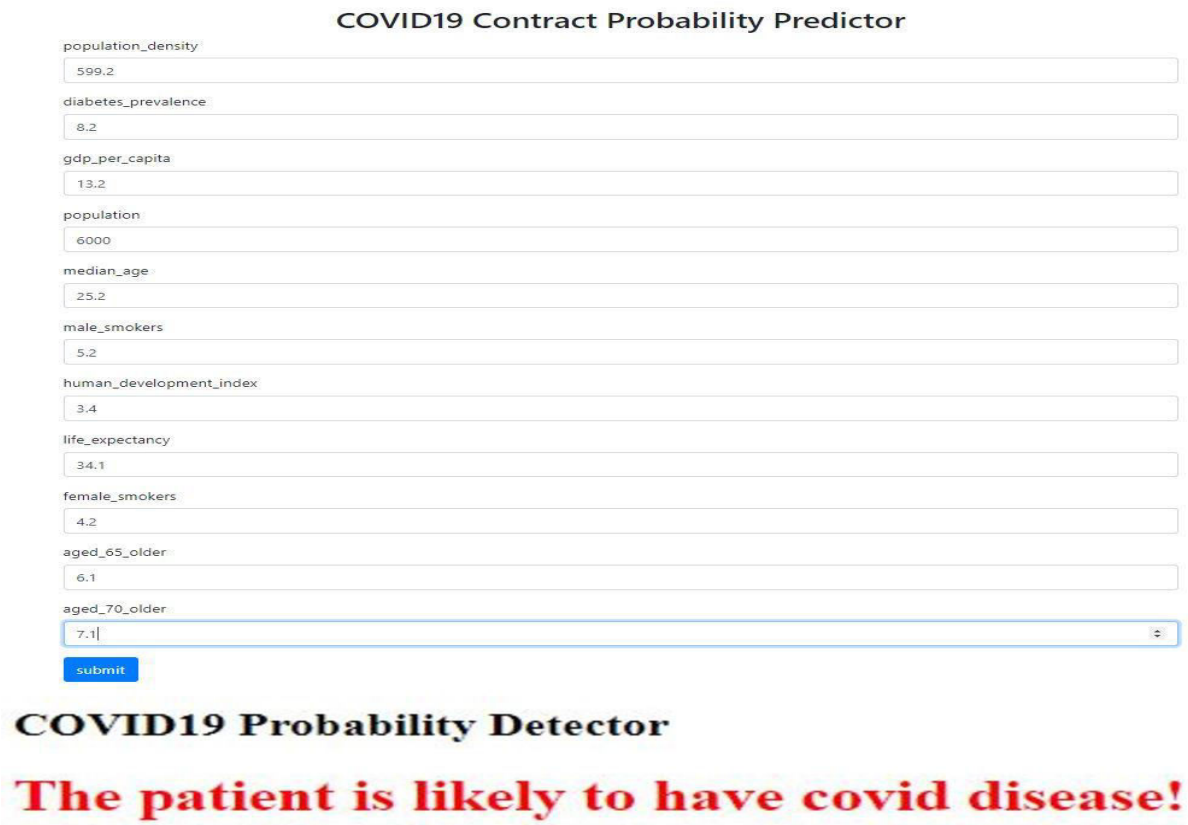
Comparison Various Approaches with our Proposed Approach

S/N	Author	Methodology	Accuracy
1	Sudirman,Ivan&Nugraha,Dimas	NB	93
2	Pourhomayoun, M., &Shakibi, M	SVM,ANN,RF,DT,LR and KNN	89.88
3	Althnian, A., Elwafa, A. A., Aloboud, N., Alrasheed, H., &Kurdi, H	DT, RF, MLP, and SVM	85.6
4	Fayyoumi, E., Idwan, S., &AboShindi, H.	LR,SVM, and MLP	91.62
5	Proposed Approach	**RF**,KNN,DT,LR,NB	94.09

Table-3

Table-3 shows the classification accuracies achieved using other methods. When compared to other methods, the proposed method achieved an accuracy of 94.09 percent. The results show that fine-tuning parameters increases classification accuracy over standard classifiers. Medical practitioners may use the tailored model to detect COVID-19 effectively.

A flask app was created giving some of the attributes as input into the Flask functions. For example, the web-application that is built using the machine learning model will be as follows,



| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 10, Issue 6, June 2021||

DOI:10.15680/IJIRSET.2021.1006022

V. CONCLUSION

The conventional and proposed systems were used to predict Novel Coronavirus disease dataset in this paper. The COVID-19 disease prediction model is generated using machine learning algorithms such as Logistic Regression, K nearest neighbour, Naive Bayes, Decision tree, and Random Forest in both cases. These models primarily consist of five phases, but the proposed model varies from the standard method in terms of hyper parameter tuning. The LR, KNN, DT, RF and NB classifiers have accuracy rates of 73.35%, 92.73%, 76.46%, 91.66% and 43.71% respectively without hyper parameters tuning. The RF, KNN, DT, LR and NB classifiers in the refined set, have accuracy rates of 94.09%, 93.89%, 89.47%, 74.18%, and 73.31% respectively using the hyper parameters tuning approach. As a result of the experimental results of performance evaluation on the COVID 19 dataset, it is concluded that the proposed model is more effective and can improve COVID-19 disease prediction. The model will be applied with a feature selection approach using various optimization strategies in the future phase of this research.

List of Abbreviations

RF	Random Forest
KNN	K-Nearest Neighbour
DT	Decision Tree
LR	Logistic Regression
NB	Naive Bayes
ML	Machine Learning
SVM	Support Vector Machine
ANN	Artificial Neural Network
MLP	Multi-Layer Perceptron

REFERENCES

[1].Bernard Stoecklin, Sibylle, et al. “First Cases of Coronavirus Disease 2019 (COVID-19) in France: Surveillance, Investigations and Control Measures, January 2020.” Eurosurveillance, vol. 25, no. 6, 2020. Crossref, doi:10.2807/1560-7917.es.2020.25.6.2000094.
[2].Github repository. 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by Johns Hopkins CSSE. Retrieved March 31, 2020 from https://github.com/CSSEGISandData/COVID-19.
[3].WHO coronaviruses (COVID-19). Retrieved March 30, 2020 from https://www.who.int/emergencies/diseases/novel-coronavirus-2019
[4]. Mantovani, Rafael G., et al. “Hyper-Parameter Tuning of a Decision Tree Induction Algorithm.” 2016 5th Brazilian Conference on Intelligent Systems (BRACIS), 2016. Crossref, doi:10.1109/bracis.2016.018.
[5].Casiraghi, Elena, et al. “Explainable Machine Learning for Early Assessment of COVID-19 Risk Prediction in Emergency Departments.” IEEE Access, vol. 8, 2020, pp. 196299–325. Crossref, doi:10.1109/access.2020.3034032.
[6].Liu, Kai, et al. “Clinical Features of COVID-19 in Elderly Patients: A Comparison with Young and Middle-Aged Patients.” Journal of Infection, vol. 80, no. 6, 2020, pp. e14–18. Crossref, doi:10.1016/j.jinf.2020.03.005.
[7].Casiraghi, Elena, et al. “Explainable Machine Learning for Early Assessment of COVID-19 Risk Prediction in Emergency Departments.” IEEE Access, vol. 8, 2020, pp. 196299–325. Crossref, doi:10.1109/access.2020.3034032.
[8].Liu, Kai, et al. “Clinical Features of COVID-19 in Elderly Patients: A Comparison with Young and Middle-Aged Patients.” Journal of Infection, vol. 80, no. 6, 2020, pp. e14–18. Crossref, doi:10.1016/j.jinf.2020.03.005.
[9].Benvenuto, Domenico, et al. “Application of the ARIMA Model on the COVID-2019 Epidemic Dataset.” Data in Brief, vol. 29, 2020, p. 105340. Crossref, doi:10.1016/j.dib.2020.105340.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 10, Issue 6, June 2021||

DOI:10.15680/IJIRSET.2021.1006022

[10].Majumdar, Manidipa, and Soudeep Deb. "A Time Series Method to Analyze Incidence Pattern and Estimate Reproduction Number of COVID-19." ArXiv, Mar. 2020, arxiv.org/pdf/2003.10655.pdf.
[11].Kucharski, Adam J., et al. "Early Dynamics of Transmission and Control of COVID-19: A Mathematical Mode lling Study." The Lancet Infectious Diseases, vol. 20, no. 5, 2020, pp. 553–58. Crossref, doi:10.1016/s1473-3099(20)30144-4.
[12].Dey, Samrat K., et al. "Analyzing the Epidemiological Outbreak of COVID-19: A Visual Exploratory Data Analysis Approach." Journal of Medical Virology, vol. 92, no. 6, 2020, pp. 632–38. Crossref, doi:10.1002/jmv.25743.
[13].Lauer, Stephen A., et al. "The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application." Annals of Internal Medicine, vol. 172, no. 9, 2020, pp. 577–82. Crossref, doi:10.7326/m20-0504.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details